ANLP Assignment 1 2021  
Marked anonymously: do not add your name(s) or ID numbers.

Use this template file for your solutions. Please start each question on a new page (as we have done here), do not remove the pagebreaks. Different questions may be marked by different markers, so your answers should be self-contained.

Don’t forget to copy your code for questions 1-5 into the Appendix of this file and to save your final version, then convert to .pdf for submission.

# Preprocessing each line (10 marks)

#We convert set to string for O(1) lookup

validCharacters = set("1234567890abcdefghijklmnopqrstuvwxyz. ")

#Process each line and add start/ending symbols

def preprocess\_line(line):

    #Add start-sentence symbol

    valid = "##"

    #Because each line ends with newline symbol, we simply delete them

    line=line.replace("\n","")

    #.lower() automatically changes all capitals to lower.

    for i in line.lower():

        if i in validCharacters:

            if i in set("1234567890"):

                valid += "0"

            else:

                valid += i

    #add end-sentence symbol

    valid+="#"

return valid

# Examining a pre-trained model (10 marks)

As every possible trigram possibility has an associated probability, and many of these trigrams have the same probability, we believe that Add-Alpha Smoothing was likely the estimation method used, of which the probability 3.333e-02 is probably the alpha if the trigram doesn’t actually exist within the corpus.

# Implementing a model: description and example probabilities (35 marks)

## Model description

Replace this text with your answer to this part of the question.

## Model excerpt

Replace this text with your answer to this part of the question.

# Generating from models (15 marks)

Replace this text with your answer to this question.

# Computing perplexity (15 marks)

Replace this text with your answer to this question.

# Extra question (15 marks)

# Appendix: your code

Include a verbatim copy of your code for questions 1-5 here. If you answered question 6, you do *not* need to include that code.

Replace this line with your code. (Make sure the spacing looks right. Ideally, use Courier or another fixed-width font, like this one. However, if you have trouble getting the code to look right, you may use screenshot images. Either way, make sure your code is legible.)